Subject: Ideographic Character Handling
Posted by royharrison on Sun, 11 Jan 2015 19:58:43 GMT
View Forum Message <> Reply to Message

I am new to U++ (but an old hand at C++). My expectation is that all programs will be written so that they will work with any character set (korean, chinese, japanese, english even). I am unsure as to how the U++ class library expects one to do this.

I see that there is a WString class but it seems to be little used by the class library itself and thus does not appear to be something you could use generally in a program. The class for looking at directories has nothing to do with it for example.

There is also a String class and this could do the job if used with UTF8 encoding. I can only assume this is the intention but, if so, I might have expected a little help (e.g. an iterator that would iterate through characters rather than bytes). It also occurs to me that maybe support for WString in the class library is in the process of being added.

I am thus left unsure as to what U++ expects me to do. This is almost certainly because I am new to U++.

Thanks in advance for any enlightenment. Roy Harrison

Subject: Re: Ideographic Character Handling
Posted by mirek on Mon, 12 Jan 2015 17:53:08 GMT
View Forum Message <> Reply to Message

royharrison wrote on Sun, 11 January 2015 20:58I am new to U++ (but an old hand at C++). My expectation is that all programs will be written so that they will work with any character set (korean, chinese, japanese, english even). I am unsure as to how the U++ class library expects one to do this.

I see that there is a WString class but it seems to be little used by the class library itself and thus does not appear to be something you could use generally in a program. The class for looking at directories has nothing to do with it for example.

There is also a String class and this could do the job if used with UTF8 encoding. I can only assume this is the intention but, if so, I might have expected a little help (e.g. an iterator that would iterate through characters rather than bytes). It also occurs to me that maybe support for WString in the class library is in the process of being added.

I am thus left unsure as to what U++ expects me to do. This is almost certainly because I am new to U++.

Thanks in advance for any enlightenment. Roy Harrison

Default way is indeed utf-8 in String.

Use conversion to WString and back for situations where you need to handle individual characters (note that there are WString::ToString/String::ToWString methods to make it easy).

The fact that you do not see many WStrings all around is in fact indication that for most time, handling of individual characters is not much needed and texts are stored as utf-8.

Iterators might sound good, but in practice you very often need random access. Random access iterator over utf-8 might be possible, but hardly any faster than simple toggling WString<->String.

(Note: It is even possible to change default charset from utf-8 to something else, like Win-1250, but it is not recommended and maintained only for backwards compatibility).

Mirek
